Mitigating Social Hazards: Early Detection of Fake News via Diffusion-Guided Propagation Path Generation

Anonymous Authors

A APPENDIX

A.1 The optimization of diffusion model

The optimization of the underlying data generating distribution $p_\theta(\mathbf{x}^0)$ is performed by optimizing the variational bound of negative log-likelihood. The objective function can be written as the KL divergence between $q\left(\mathbf{x}^{0:T}\right)$ and $p_\theta\left(\mathbf{x}^{0:T}\right)$:

$$\begin{aligned}
&\mathbb{E}\left[-\log p_\theta(\mathbf{x}^0)\right] \\
&\leq D_{KL}\left(q\left(\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_T\right) \| p_\theta\left(\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_T\right)\right) \\
&= \mathbb{E}_q\left[-\log p\left(\mathbf{x}_T\right) - \sum_{t=1}^{T} \log \frac{p_\theta\left(\mathbf{x}_{t-1} \mid \mathbf{x}_t\right)}{q\left(\mathbf{x}_t \mid \mathbf{x}_{t-1}\right)}\right] + C_1 \\
&= \sum_{t=1}^{T} \underbrace{D_{KL}\left(q(\mathbf{x}^{t-1}|\mathbf{x}^t, \mathbf{x}^0)||p_\theta(\mathbf{x}^{t-1}|\mathbf{x}^t)\right)}_{:=L_{t-1}} + C_2,
\end{aligned} \tag{1}$$

where C_1 and C_2 are constants that are independent of the model parameter θ. Using Bayes' theorem, the posterior distribution $q(\mathbf{x}^{t-1}|\mathbf{x}^t, \mathbf{x}^0)$ could be solved in closed form:

$$q\left(\mathbf{x}^{t-1} \mid \mathbf{x}^t, \mathbf{x}^0\right) = \mathcal{N}\left(\mathbf{x}^{t-1}; \tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}^t, \mathbf{x}^0\right), \tilde{\beta}_t \mathbf{I}\right), \tag{2}$$

where

$$\begin{aligned}
\tilde{\boldsymbol{\mu}}_t\left(\mathbf{x}^t, \mathbf{x}^0\right) &= \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}^0 + \frac{\sqrt{\alpha_t}\left(1-\bar{\alpha}_t\right)}{1-\bar{\alpha}_t}\mathbf{x}^t \quad , \\
\tilde{\beta}_t &= \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t.
\end{aligned} \tag{3}$$

Further reparameterizing $\boldsymbol{\mu}_\theta\left(\mathbf{x}^t, t\right)$ as:

$$\boldsymbol{\mu}_\theta\left(\mathbf{x}^t, t\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}^t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\theta\left(\mathbf{x}^t, t\right)\right). \tag{4}$$

The training objective in Eq. 1 is simplified as:

$$\begin{aligned}
L =& \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}||\tilde{\mu}_t(\mathbf{x}_t, \mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t, t)||^2\right] \\
=& \mathbb{E}_{\mathbf{x}^0, \boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\left(\sqrt{\bar{\alpha}_t}\mathbf{x}^0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t\right)\right\|^2\right].
\end{aligned} \tag{5}$$

A.2 Datasets Details.

We conduct experiments on three real-world datasets, Pheme[25], PolitiFact and GossipCop[28]. Pheme is collected from Twitter, a widely used social media platform. PolitiFact and GossipCop are collected from fact-check websites which are created to verify the authenticity of published news. We conduct data cleaning on all three datasets and Table 1 presents the basic information.

A.3 Baselines.

We compare our model in two categories of baselines, a total of ten models.

The content-based fake news detection models:

- **BERT** [7] is a pretrained model to extract text features.

Dataset	Pheme	PolitiFact	GossipCop
#. total news	3,506	484	5,526
#. fake news	1,142	297	1,935
#. real news	2,364	187	3,591
#. words per news	13.8	1546.1	578.6

Table 1: Statistics of the datasets

- **MVAE** [15] comprises an encoder, a decoder, and a detector to classify fake news.
- **EANN** [34] designs a multi-task learning framework to detect fake news and classify events simultaneously.
- **CAFE** [6] aggregates unimodal features and cross-modal correlations to help detection.
- **LLAMA2** [32] applies a straightforward prompt engineering such as "Is it true that x? Yes or no?", where 'x' represents news text content.
- **Detect-GPT** [2] develops a zero-shot prompt engineering for text content fake news detection.

The propagation-enhanced detection methods:

- **CSI** [27] employs LSTM to encode the news content, and utilizes the group behavior of users for detection.
- **Bi-GCN** [1] uses a Bidirectional Graph Convolutional Network to learn the propagation patterns of misinformation.
- **UPFD** [8] learns user preferences through their past engaged posts, and combines content with graph modeling.
- **MFAN** [43] integrates textual, visual, and social graph features in one unified framework for rumor detection.

A.4 Implementation Details.

We employ AdamW [18] as the optimizer, and the batch size is set at 16. The initial learning rate is set to $5e^{-3}$. The hidden size of user embedding is set to 128. The selection parameter Top-K is set as 5. The unconditional training probability λ as 0.1 [10]. The total diffusion step T is searched in the range of [100, 200, 300, 400, 500, 600], while the conditional guidance strength w is in the range of [1,2,3,4,5,6,7,8].

A.5 Related work

Content-based. In recent years, there has been widespread attention on the automated detection of fake news in social media. Some early research endeavors seek to detect fake news by extracting features from the textual content of the news articles [4, 9, 14, 19, 22]. For instance, Wawer et al. [35] utilizes textual and linguistic features from websites, based on bag-of-words vector space and psycholinguistic dimensions, to predict the credibility of websites. Vaibhav et al. [33] proposes a model based on graph neural networks to capture the interaction among sentences in the content of fake news. Meanwhile, some researchers propose novel approaches utilizing cross-modal features in news to enhance the accuracy of fake news detection [6, 13, 30, 34, 36]. Khattar et al. [15] proposes the MVAE model which consists of three components: an encoder, a decoder, and a fake news classifier. The encoder is used to encode a shared representation of features, the decoder reconstructs the data from multi-modal representations, and the fake news classifier categorizes news into true or false categories. Zhou et al. [44] proposes the SAFE method by computing the correlation between textual information and visual information, defining it as a modified cosine similarity to detect fake news. Wu et al. [39] uses multiple co-attention layers to learn the relationship between text and images.

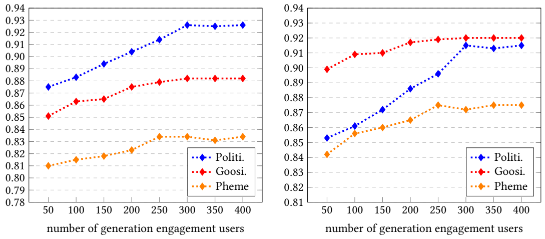
Recently, researchers start exploring the utilization of Large Language Models (LLMs) for fake news detection. Some studies focus on directly prompting various LLMs such as GPT-3 [2], ChatGPT-3.5 [3, 11] and GPT-4 [5] for misinformation detection. For instance, Chen et al. [5] investigates ChatGPT-3.5 and GPT-4 using both standard prompting strategies and zero-shot chain-of-thought prompting strategies for detecting human-written misinformation. Pan et al. [24] introduces a program-guided fact-checking framework leveraging the contextual learning ability of LLMs to generate reasoning programs guiding veracity verification. Wu et al. [37] applies GPT-3.5 as a feature extractor to identify out-of-context images. However, existing LLM-based fake news detection methods primarily rely on textual semantics, often insufficient for effectively considering user behaviors during news dissemination. Textual features alone may not adequately verify the veracity of news items in certain situations.

Propagation-enhanced. Different from methods that rely on the content of news for fake news detection, Propagation Graph-enhanced fake news detection approaches aim to improve accuracy by leveraging differences in the propagation processes between real and fake news [20, 26, 29, 38, 45]. Jin et al. [12] applies epidemiological models to characterize information cascades triggered by both real and fake news on Twitter. Wu et al. [38] proposes a graph kernel-based SVM classifier to detect fake news by learning high-order propagation patterns. Ma et al. [21]designes a model based on Recursive Neural Network (RNN) to represent features of news by integrating both the propagation structure and content features of the news. Zhang et al [41]. proposes a deep diffusive Network model that can simultaneously learn latent representations and infer the accuracy of news articles, creators, and topics. Ma et al. [20] proposes a graph kernel-based SVM classifier that captures high-order patterns distinguishing different types of fake news by evaluating the similarity between their propagation tree structures. Liu et al. [16] conducts authenticity assessment of news based on user profile information within the news propagation network

Due to the potential impact of misinformation on a large audience and its negative consequences during dissemination, the early detection of fake news has become a crucial research focus within the field of fake news detection. Zhao et al.[42] posit that false information is more likely to arouse user suspicion. They propose aggregating relevant articles using specific phrases and subsequently employing a cluster-based classifier for early detection of false news propagation.Yang et al. [40] endeavor to utilize convolutional neural networks for extracting linguistic and user features from news content and employ these features for the early detection of fake news. Nguyen et al.[23] emploies deep neural networks to automatically capture features at a posterior level, achieving superior performance in fake news early detection. Liu et al. [17] proposes a novel deep neural network that integrates crowd response features and user reactions, effectively enabling early detection of misinformation. Song et al. [31] introduces the concept of trust checkpoints, suggesting collecting every 10 posts along the timeline as a time step for the Recurrent Neural Network (RNN) and making predictions at each step.

A.6 Analysis of Generation User Number

We evaluate detection performance on the various number of generation users in Figure 7. With the increase of generation users, the model performance first increases and then levels off. This demonstrates that the performance does not increase after a certain amount of propagated information.



(a) Accuracy on Diffusion (b) F1-score on Diffusion

Figure 1: Performance with different generation users.

A.7 Complexity Analysis

The time complexity of diffusion guided generation is $O(T)$, where T is the times of diffusion step. The time complexity of directed propagation graph is $O(|U^{\mathcal{H}}|^2)$ and user hypergraph is $O(|U^G|^2)$. The space complexity of the whole model is $O(N^2|U|^2)$, where N is the hidden size of user embedding.

REFERENCES

[1] Tian Bian, Xi Xiao, Tingyang Xu, Peilin Zhao, Wenbing Huang, Yu Rong, and Junzhou Huang. 2020. Rumor detection on social media with bi-directional graph convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 549–556.

[2] Mars Gokturk Buchholz. 2023. Assessing the effectiveness of gpt-3 in detecting false political statements: A case study on the liar dataset. *arXiv preprint arXiv:2306.08190* (2023).

[3] Kevin Matthe Caramancion. 2023. Harnessing the power of ChatGPT to decimate mis/disinformation: Using ChatGPT for fake news detection. In *2023 IEEE World AI IoT Congress (AIIoT)*. IEEE, 0042–0046.

[4] Carlos Castillo, Marcelo Mendoza, and Barbara Poblete. 2011. Information credibility on twitter. In *Proceedings of the 20th international conference on World wide web*. 675–684.

[5] Canyu Chen and Kai Shu. 2023. Can llm-generated misinformation be detected? *arXiv preprint arXiv:2309.13788* (2023).

[6] Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. 2022. Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022*. 2897–2905.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[8] Yingtong Dou, Kai Shu, Congying Xia, Philip S Yu, and Lichao Sun. 2021. User preference-aware fake news detection. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2051–2055.

[9] Aditi Gupta, Hemank Lamba, Ponnurangam Kumaraguru, and Anupam Joshi. 2013. Faking sandy: characterizing and identifying fake images on twitter during hurricane sandy. In *Proceedings of the 22nd international conference on World Wide Web*. 729–736.

[10] Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

[11] Yue Huang and Lichao Sun. 2023. Harnessing the power of chatgpt in fake news: An in-depth exploration in generation, detection and explanation. *arXiv preprint arXiv:2310.05046* (2023).

[12] Fang Jin, Edward Dougherty, Parang Saraf, Yang Cao, and Naren Ramakrishnan. 2013. Epidemiological modeling of news and rumors on twitter. In *Proceedings of the 7th workshop on social network mining and analysis*. 1–9.

[13] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM international conference on Multimedia*. 795–816.

[14] Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. 2016. Novel visual and statistical image features for microblogs news verification. *IEEE transactions on multimedia* 19, 3 (2016), 598–608.

[15] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. Mvae: Multimodal variational autoencoder for fake news detection. In *The world wide web conference*. 2915–2921.

[16] Yang Liu and Yi-Fang Wu. 2018. Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 32.

[17] Yang Liu and Yi-Fang Brook Wu. 2020. Fned: a deep network for fake news early detection on social media. *ACM Transactions on Information Systems (TOIS)* 38, 3 (2020), 1–33.

[18] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[19] Jing Ma, Wei Gao, Prasenjit Mitra, Sejeong Kwon, Bernard J Jansen, Kam-Fai Wong, and Meeyoung Cha. 2016. Detecting rumors from microblogs with recurrent neural networks. (2016).

[20] Jing Ma, Wei Gao, and Kam-Fai Wong. 2017. Detect rumors in microblog posts using propagation structure via kernel learning. Association for Computational Linguistics.

[21] Jing Ma, Wei Gao, and Kam-Fai Wong. 2018. Rumor detection on twitter with tree-structured recursive neural networks. Association for Computational Linguistics.

[22] Jing Ma, Wei Gao, and Kam-Fai Wong. 2019. Detect rumors on twitter by promoting information campaigns with generative adversarial learning. In *The world wide web conference*. 3049–3055.

[23] Tu Ngoc Nguyen, Cheng Li, and Claudia Niederée. 2017. On early-stage debunking rumors on twitter: Leveraging the wisdom of weak learners. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II 9*. Springer, 141–158.

[24] Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. *arXiv preprint arXiv:2305.12744* (2023).

[25] Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. 2019. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE international conference on data mining (ICDM)*. IEEE, 518–527.

[26] Nir Rosenfeld, Aron Szanto, and David C Parkes. 2020. A kernel of truth: Determining rumor veracity on twitter by diffusion pattern alone. In *Proceedings of The Web Conference 2020*. 1018–1028.

[27] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 797–806.

[28] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data* 8, 3 (2020), 171–188.

[29] Kai Shu, Deepak Mahudeswaran, Suhang Wang, and Huan Liu. 2020. Hierarchical propagation networks for fake news detection: Investigation and exploitation. In *Proceedings of the international AAAI conference on web and social media*, Vol. 14. 626–637.

[30] Shivangi Singhal, Anubha Kabra, Mohit Sharma, Rajiv Ratn Shah, Tanmoy Chakraborty, and Ponnurangam Kumaraguru. 2020. Spotfake+: A multimodal framework for fake news detection via transfer learning (student abstract). In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 34. 13915–13916.

[31] Changhe Song, Cheng Yang, Huimin Chen, Cunchao Tu, Zhiyuan Liu, and Maosong Sun. 2019. CED: credible early detection of social media rumors. *IEEE Transactions on Knowledge and Data Engineering* 33, 8 (2019), 3035–3047.

[32] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023).

[33] Raghuram Mandyam Annasamy Vaibhav and Eduard Hovy. 2019. Do Sentence Interactions Matter? Leveraging Sentence Level Representations for Fake News Classification. *EMNLP-IJCNLP 2019* (2019), 134.

[34] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. Eann: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. 849–857.

[35] Aleksander Wawer, Radoslaw Nielek, and Adam Wierzbicki. 2014. Predicting webpage credibility using linguistic features. In *Proceedings of the 23rd international conference on world wide web*. 1135–1140.

[36] Zimian Wei, Hengyue Pan, Linbo Qiao, Xin Niu, Peijie Dong, and Dongsheng Li. 2022. Cross-modal knowledge distillation in multi-modal fake news detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4733–4737.

[37] Guangyang Wu, Weijie Wu, Xiaohong Liu, Kele Xu, Tianjiao Wan, and Wenyi Wang. 2023. Cheap-fake Detection with LLM using Prompt Engineering. In *2023 IEEE International Conference on Multimedia and Expo Workshops (ICMEW)*. IEEE, 105–109.

[38] Ke Wu, Song Yang, and Kenny Q Zhu. 2015. False rumors detection on sina weibo by propagation structures. In *2015 IEEE 31st international conference on data engineering*. IEEE, 651–662.

[39] Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, and Zhen Xu. 2021. Multimodal fusion with co-attention networks for fake news detection. In *Findings of the association for computational linguistics: ACL-IJCNLP 2021*. 2560–2569.

[40] Yang Yang, Lei Zheng, Jiawei Zhang, Qingcai Cui, Zhoujun Li, and Philip S Yu. 2018. TI-CNN: Convolutional neural networks for fake news detection. *arXiv preprint arXiv:1806.00749* (2018).

[41] Jiawei Zhang, Bowen Dong, and S Yu Philip. 2020. Fakedetector: Effective fake news detection with deep diffusive neural network. In *2020 IEEE 36th international conference on data engineering (ICDE)*. IEEE, 1826–1829.

[42] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th international conference on world wide web*. 1395–1405.

[43] Jiaqi Zheng, Xi Zhang, Sanchuan Guo, Quan Wang, Wenyu Zang, and Yongdong Zhang. 2022. MFAN: Multi-modal Feature-enhanced Attention Networks for Rumor Detection. IJCAI.

[44] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. SAFE: Similarity-Aware Multi-modal Fake News Detection. In *Pacific-Asia Conference on knowledge discovery and data mining*. Springer, 354–367.

[45] Xinyi Zhou and Reza Zafarani. 2019. Network-based fake news detection: A pattern-driven approach. *ACM SIGKDD explorations newsletter* 21, 2 (2019), 48–60.